ICS 35.080
CCS L 76

DB14

山 西 省 地 方 标 准

DB14/T 2527—2022

云平台 人工智能建模系统框架及功能要求

2022-08-18 发布　　2022-11-18 实施

山西省市场监督管理局　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由山西省工业和信息化厅提出、组织实施和监督检查。

山西省市场监督管理局对文件的组织实施情况进行监督检查。

本文件由山西省网络安全和大数据信息技术标准化技术委员会归口。

本文件起草单位：山西云时代技术有限公司、山西省信息产业技术研究院有限公司、山西云时代政务云技术有限公司、山西数字政府建设运营有限公司。

本文件主要起草人：盛佃清、王文逾、侯彦英、郝俊宇、康晓丽、刘宁、申利华、李华、郑亮、王奇侠、白鹏、郭靖伟、李潞洋、赵世琛、杜军军、吕云云、张弋、杨峰光、杜亮亮、王忠民、李雨萌、孙凯凯、胡博、崔志学、肖晋飞、温静、高俊杰、付玉辉、徐流明、许兴欣、董力熇、张培玉、田垒、郑立、韩思齐。

云平台人工智能建模系统功能要求

1 范围

本文件规定了云平台人工智能建模系统的各组件功能要求。

本文件适用于云平台上人工智能建模系统及解决方案的数据处理、算法设计、模型训练、模型管理等功能要求，可作为云平台上人工智能建模系统的规划、设计、建设、评估及验收的依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.31-2006 信息技术 词汇 第31部分：人工智能 机器学习

GB/T 5271.34-2006 信息技术 词汇 第34部分：人工智能 神经网络

3 术语和定义

GB/T 5271.31-2006，GB/T 5271.34-2006界定的以及下列术语和定义适用于本文件。

3.1

云平台

本文件所指云平台是面向全省域，为政府、社会团体和企事业组织提供专业化服务的一体化云服务体系。

3.2

人工智能建模系统

为数据分析人员、业务建模人员和模型管理人员提供数据处理、模型构建与训练、模型部署与管理解决方案的模型平台。

3.3

算子

构成人工智能建模算法的计算单元。

3.4

特征工程

利用领域知识从原始数据中提取特征的过程。

4 缩略语

下列缩略语适用于本文件。

AI：人工智能（Artificial Intelligence）

DAG：有向无环图（Directed Acyclic Graph）

NLP：自然语言处理（Natural Language Processing）

JDBC：Java数据库连接（Java Database Connectivity）

HDFS：Hadoop分布式文件系统（Hadoop Distributed File System）

SQL：结构化查询语言（Structured Query Language）

API：应用程序接口（Application Programming Interface）

ROC：接收者操作特征（Receiver Operating Characteristic）

PR：查全率（Precision-Recall）

REST： 表述性状态转移 （Representational State Transfer）

5 功能要求

5.1 概述

云平台人工智能建模系统的功能框架见图1，包括数据导入导出、数据预览与探索、数据预处理、特征工程、算法选择、模型训练与评估、模型管理、模型市场、工作流调度等核心能力。

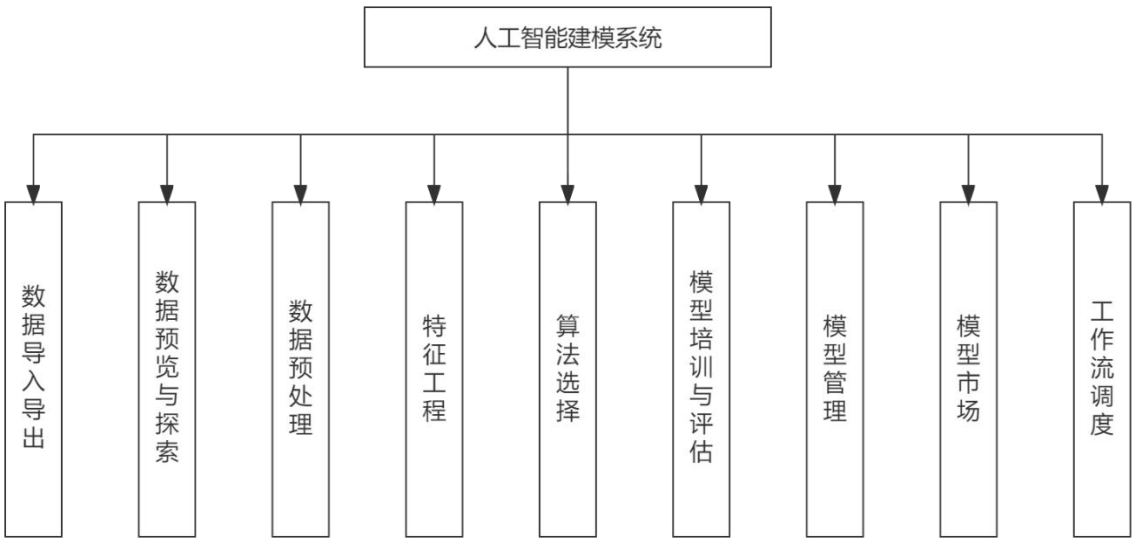

图 1 云平台人工智能建模系统的功能框架

5.2 数据导入导出

5.2.1 数据导入

支持多种数据源包括关系型数据库、Hive、HBase、ElasticSearch、HDFS、文件格式、JDBC等，同时支持数据导入时转换数据类型。数据源接入使用统一视图及规范。

5.2.2 数据导出

支持将结果数据导出至关系型数据库、Hive、HDFS、JDBC等，同时支持结果数据导出至数据源。

5.2.3 数据样例

人工智能平台应提供不同类型的样例数据以供测试。

5.3 数据预览与探索

5.3.1 数据质量分析

支持对脏数据，数据缺失值、异常值等的检查。

5.3.2 数据统计分析

支持查看数据的分布情况和统计学指标。支持图形化自定义统计分析数据。

5.3.3 数据特征分析

支持在数据集合进行分布分析，对比分析，统计量分析和相关分析，为数据建模人员提供基本的特征描述。

5.3.4 复杂数据特征分析

支持交互式分析和探索的编程环境。包括R、Python等编程环境，用于复杂的数据特征分析。

5.4 数据预处理

5.4.1 数据清洗

支持按照预定义的清洗模式对全量数据进行原始无效异常数据过滤和缺失数据补齐。

5.4.2 数据变换

提供包括数据属性转换、新属性生成在内的处理能力。

5.4.3 数据规约

提供对基本数据属性的归一化处理能力。

5.4.4 自动化预处理

支持数据预处理自动化，包括自动填充、自动清理、自动转换以及自动归一化等。

5.4.5 预处理行业模板

人工智能平台应提供预处理操作算子样例及常用模板。

5.5 特征工程

5.5.1 特征工程流程

特征工程流程包括特征变换、特征重要性评估、特征选择、特征生成等。

5.5.2 特征工程自动化

特征工程自动化包括自动多表扩展、自动特征变换、自动特征选择以及自动特征生成等。

5.5.3 特征提取模板

支持特征提取算子和模板配置。

5.6 算法选择

5.6.1 基础能力

支持多种优化算法，算法参数可配置。

5.6.2 支持但不限于以下的算法类型

特征权重、流处理、预处理、表操作、机器学习、图嵌入、验证与评估、NLP、时间序列、统计、集成学习、深度学习、图计算、图像处理、强化学习等。

5.6.3 自定义算法

支持通过Python，R等实现自定义算法，支持用户自定义持久化扩展算子库。

5.6.4 实用工具库

提供支持子流程、添加宏、提取宏、生成宏、删除宏、循环，支持子流程的自定义封装和命名，支持自定义单机脚本算子快速实现分布式化等功能的实用工具。

5.6.5 算法样例库

提供章节5.6.2、5.6.3所列算法的使用样例。

5.7 模型训练与评估

5.7.1 训练过程

可以启动和停止训练任务，可以查看运行日志。训练过程中支持调试功能，可进行单步调试，断点调试。支持训练过程中间数据查看、导出。

5.7.2 资源共享

支持多个用户分组管理和共享计算资源。

5.7.3 资源管控

支持对物理资源进行虚拟化管控，可以动态进行资源的申请或释放。

5.7.4 复杂任务依赖

支持多任务之间图形化构建依赖，以构建复杂的模型训练任务及数据分析任务。

5.7.5 自动调参与自动建模

支持自动调参和搜索网格，包括在给定命中率和覆盖率的要求下搜索参数输出结果，及在给定参数下搜索最优结果。

支持自动建模，自动选择算法及参数。

5.7.6 交叉验证

支持按比例随机分配训练与测试集，支持交叉检验。

5.7.7 评估指标

支持多种评估指标，如混淆矩阵，ROC曲线，PR曲线，加权召回率等。对于二分类，输出包括评价指标的数目表格；对于多分类，输出混淆矩阵。

5.7.8 评估样例库

提供所有评估算子样例。

5.8 模型管理

5.8.1 模型的版本管理

支持历史、新建及外部导入模型的保存和版本管理，支持模型详细查看，模型结果查看。

5.8.2 模型导入导出

支持多种模型格式。支持导出Json模型，包括聚类、分类、回归等类型。

5.8.3 深度学习模型管理

支持深度学习模型导入导出和可视化查看，支持实验应用。

5.9 模型市场

5.9.1 模型用户管理

支持管理员对其所属普通用户项目情况及权限进行管理。

5.9.2 模型服务上架

支持任务/实验、代码、自定义镜像等在模型市场上架。

5.9.3 模型服务上、下线

支持模型服务的上、下线与列表查看。

5.9.4 模型服务更新

支持滚动更新及灰度更新，且灰度升级支持分配流量权重。

5.9.5 模型服务测试

支持服务上线后的API测试。

5.9.6 模型服务管理

支持自定义模型部署,生成相应REST API，手动增加实例数量提高服务的负载均衡；可查看当前导入平台的API列表。

5.9.7 模型服务监控

支持线上模型服务监控，可查看模型服务内容、运行状态、实例详情、资源设置等，后台可以统计API的调用情况和结果统计。

5.9.8 模型服务使用

API服务上线后，可通过REST API调用，传入参数并获得预测值。

5.10 工作流调度

5.10.1 任务配置

支持可视化建模、代码建模、特征和模型上架、上线等任务类型。支持对单个任务进行资源配置，如可视化建模、代码建模等。

5.10.2 设计工作流

任务定义成功后，确定各任务彼此间的逻辑依赖关系，任务会自上而下执行。支持通过Cron表达式，来设置整个工作流的调度周期。

5.10.3 执行工作流

支持对工作流进行调试，确保整体流程可执行，再进行调度。设置工作流的调度周期后，工作流会按照设置的周期定时调度。

5.10.4 工作流上、下线

支持对工作流进行上线、下线操作。

5.10.5 工作流导入导出

支持从外部导入工作流，支持工作流导出到本地，导入导出文件为JSON格式。

5.10.6 工作流详情

支持查看单个工作流每次的执行时间和执行状态。支持查看工作流下的单个任务每次的执行时间、状态和日志详情。

参 考 文 献

[1] 《国家新一代人工智能标准体系建设指南》国标委联〔2020〕35